Supplementary Material for TokenMatcher: Diverse Tokens Matching for Unsupervised Visible-Infrared Person Re-Identification

Infrared To Visible Learning

The $\mathcal{L}_{\text{R2V}}$ can be formulated as follows:

$$\phi_k^{rs} = \frac{1}{|\mathcal{H}_k^r \bigcup \mathcal{H}_{\text{R2V}[k]}^r|}\Big(\sum_{u_n^r \in \mathcal{H}_k^r} u_n^r + \sum_{u_n^v \in \mathcal{H}_{\text{R2V}[k][0]}^v} u_n^v\Big), \quad (1)$$

$$\mathcal{L}_{\text{R2V}} = -\log \frac{\exp\left(q_i^r \cdot \phi_{\hat{y_i^r}}^{vs}/\tau\right)}{\sum_{k=0}^{N^v} \exp\left(q_i^r \cdot \phi_k^{vs}/\tau\right)}, \quad (2)$$

where ϕ_k^{rs} is the modality-shared memory of infrared modality. Note that $\hat{y_i^r} = \text{R2V}[y_i^r][0]$. Infrared-to-visible learning allows each infrared cluster to align only with the first, i.e. most similar, visible cluster it is associated with.

Datasets

We evaluate the proposed methods on two widely-used visible infrared person re-identification, namely, SYSU-MM01 and RegDB .SYSU-MM01 contains 395 training identities from four RGB cameras and two IR cameras, comprising 22,258 RGB images and 11,909 IR images. Following existing works (Ye et al. 2020, 2021), we adopt two all-search and indoor-search test modes. The RegDB data set is collected by two aligned visible and thermal imaging cameras and contains 412 identities. Following previous works (Ye et al. 2020, 2021), we perform ten trials for gallery set selection and calculate the average performance in both test modes, i.e., thermal to visible and visible to thermal.

Implementation Details.

Our proposed method is implemented using PyTorch. Following SDCL (Yang, Chen, and Ye 2024), we adopt the feature extractor from TransReID (He et al. 2021) as the backbone network with augmented dual-contrastive learning (Yang et al. 2022), and the patch embedding layers are constructed with IBN and CNN module (Luo et al. 2021). Visible and infrared images are resized to 384 × 128 before entering the network. Channel augmentation (Ye et al. 2023), random cropping, level flipping, random erasing, and color dithering are used for data enhancement. The model is trained for a total of 100 epochs. In the first 50 epochs, we employ the baseline for learning. The proposed framework is added to the training in the last 50 epochs. The maximum distance for DBSCAN is set to 0.6 on SYSU-MM01 and 0.2 on RegDB. We use the SGD optimizer to train the model and use the cosine annealing to decrease the learning rate. For the first 50 epochs of the SYSU-MM01, the learning rate is initialized to $3.5e$-4. Beyond that, the learning rate is initialized to $3.5e$-5. The momentum factor is set to 0.2 in the baseline and homogeneous fusion module and set to 0.8 in shared memory learning. The γ is set to 0.9 following (Yang, Chen, and Ye 2024). The β_1, β_2, β_3 is set to 0.4, 0.5, 0.03 on SYSU-MM01 and 0.4, 0.1, 0.01 on RegDB. The λ in Algorithm 1 is set to 2. The number of class tokens N is set to 4 and 6 on SYSU-MM01 and RegDB. The other contrastive learning settings follow (Yang et al. 2022).

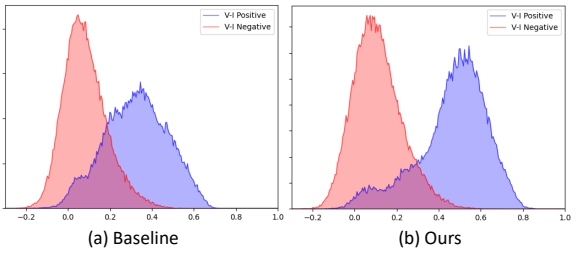

Figure 1: The similarity distribution visualization of 10 randomly selected identities.

Visualization analysis

We visualized the similarity distribution visualization and top-10 retrieved results of some example queries, as shown in Figure 1 and Figure 2. We observed that TokenMatcher demonstrates a strong ability to accurately identify individuals from a cross-modality gallery, even in scenarios where some images are challenging for human recognition due to variations in lighting, angle, or quality. This showcases the robustness of our model in handling complex and diverse data. However, the system encounters difficulties when distinguishing between pedestrians who have very similar clothing and posture, as these subtle differences are not easily discernible. This limitation underscores the need for further enhancements in our approach to better differentiate be-

tween individuals who closely resemble each other in future iterations of our work.

Limitations and Future Research. Although our approach demonstrates impressive performance, it has two limitations: 1) There is still significant room for improvement compared to supervised VI-ReID. 2) For some occluded pedestrians, it remains challenging to assign accurate pseudo-labels and cross-modal labels. In the future, it will be necessary to explore richer representations of pedestrian features, as well as effective methods for refining pseudo-labels and cross-modal labels, to provide more accurate supervision for cross-modal learning.

References

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15013–15022.

Luo, H.; Wang, P.; Xu, Y.; Ding, F.; Zhou, Y.; Wang, F.; Li, H.; and Jin, R. 2021. Self-supervised pre-training for transformer-based person re-identification. *arXiv preprint arXiv:2111.12084*.

Nguyen, D. T.; Hong, H. G.; Kim, K. W.; and Park, K. R. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3): 605.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. RGB-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, 5380–5389.

Yang, B.; Chen, J.; and Ye, M. 2024. Shallow-Deep Collaborative Learning for Unsupervised Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16870–16879.

Yang, B.; Ye, M.; Chen, J.; and Wu, Z. 2022. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2843–2851.

Ye, M.; Shen, J.; J. Crandall, D.; Shao, L.; and Luo, J. 2020. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, 229–247. Springer.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 2872–2893.

Ye, M.; Wu, Z.; Chen, C.; and Du, B. 2023. Channel augmentation for visible-infrared re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

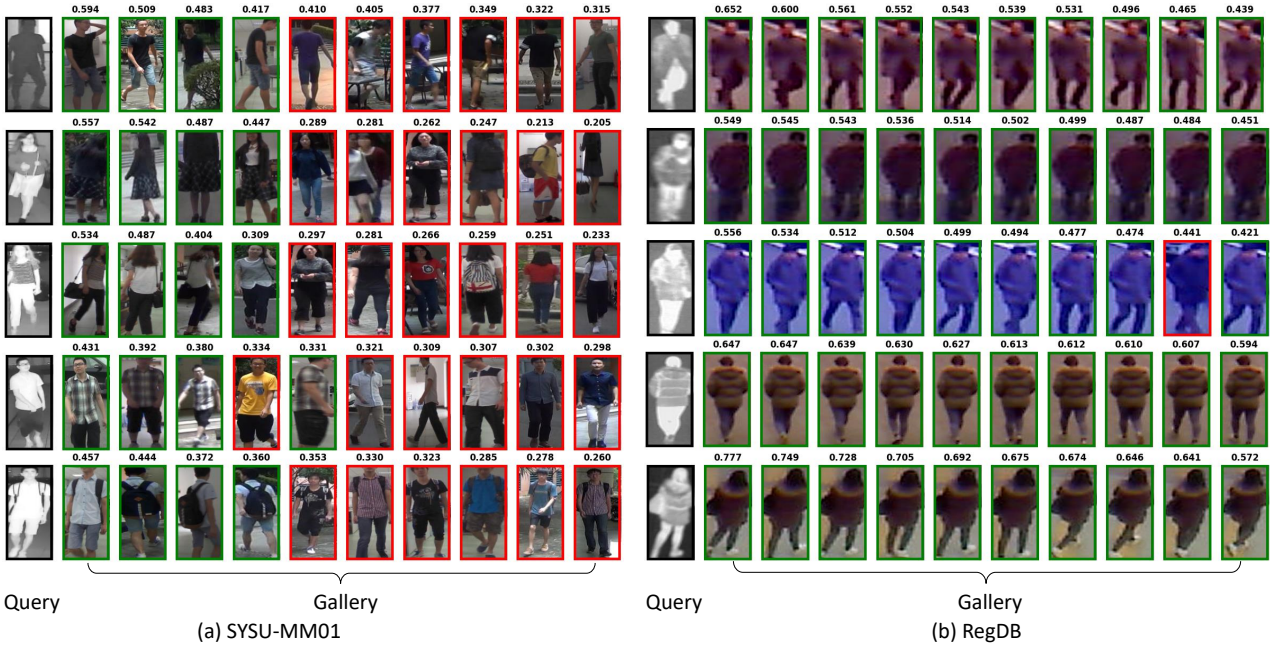

Figure 2: The top-10 retrieved results of some example queries on the SYSU-MM01 and RegDB datasets with the setting of infrared to visible. The cosine similarity score is reported for each image pair. The green bounding boxes indicate the correct matchings and the red bounding boxes represent the wrong matchings.